

# Generalized Statistical Tests for mRNA and Protein Subcellular Spatial Patterning against Complete Spatial Randomness (Preprint)

Jonathan H. Warrell<sup>1,2\*</sup>, Anca F. Savulescu<sup>1,2</sup>, Robyn Brackin<sup>1</sup>, Musa M. Mhlana<sup>1,2,3</sup>

<sup>1</sup> *Gene Expression and Biophysics group, Council for Scientific and Industrial Research, Pretoria, South Africa*

<sup>2</sup> *Division of Chemical Systems and Synthetic Biology, Faculty of Health Sciences,  
University of Cape Town, South Africa*

<sup>3</sup> *Unidade de Biofísica e Expressão Genética, Instituto de Medicina Molecular, Universidade de Lisboa, Portugal*

\*jonathan.warrell@gmail.com

**Abstract.** We derive generalized estimators for a number of spatial statistics that have been used in the analysis of spatially resolved omics data, such as Ripley's  $K$ ,  $H$  and  $L$  functions, clustering index, and degree of clustering, which allow these statistics to be calculated on data modelled by arbitrary random measures. Our estimators generalize those typically used to calculate these statistics on point process data, allowing them to be calculated on random measures which assign continuous values to spatial regions, for instance to model protein intensity. The clustering index ( $H^*$ ) compares Ripley's  $H$  function calculated empirically to its distribution under complete spatial randomness (CSR), leading us to consider CSR null hypotheses for random measures which are not point-processes when generalizing this statistic. For this purpose, we consider restricted classes of completely random measures which can be simulated directly (Gamma processes and Marked Poisson Processes), as well as the general class of all CSR random measures, for which we derive an exact permutation-test based  $H^*$  estimator. We establish several properties of the estimators we propose, including bounds on the accuracy of our general Ripley  $K$  estimator, its relationship to a previous estimator for the cross-correlation measure, and the relationship of our generalized  $H^*$  estimator to a number of previous statistics. We test the ability of our approach to identify spatial patterning on synthetic and biological data. With respect to the latter, we demonstrate our approach on mixed omics data, by using Fluorescent In Situ Hybridization (FISH) and Immunofluorescence (IF) data to probe for mRNA and protein subcellular localization patterns respectively in polarizing mouse fibroblasts on micropatterned cells. Using the generalized clustering index and degree of clustering statistics we propose, we observe correlated patterns of clustering over time for corresponding mRNAs and proteins, suggesting a deterministic effect of mRNA localization on protein localization for several pairs tested, including one case in which spatial patterning at the mRNA level has not been previously demonstrated.

## 1 Introduction

Detection of spatial patterning is important in many domains, including molecular biology [1], ecology [2] and epidemiology [3]. Spatial patterning can be identified by testing whether observed data departs from a model of spatial randomness: For instance, the homogeneous Poisson process may serve as a model of spatial randomness for point process data, and deviations from Poisson statistics may be used to detect spatial structure such as clustering. The Ripley  $K$  function is widely employed, along with associated  $L$  and  $H$  functions, to analyse deviations from homogeneous Poisson statistics [4], since it permits tests for clustering and dispersion at multiple scales. A number of related statistics have been introduced based on the  $K$  function to summarize such deviations (employing either simulations or analytic approaches to evaluate the critical quantiles under Poisson statistics), including the clustering index and degree of clustering [1], and a variance normalized alternative to the  $L$  function [5]. The above spatial statistics have been defined and applied in the context of point process, where the data to be analysed consists of a collection of points in (typically) Euclidean 2- or 3-space: For instance, the degree of clustering has been applied to the study of the spatial distribution of individual mRNA transcripts from a single gene, treated as point particles at positions inferred from Fluorescent In Situ Hybridization (FISH) microscopy data, whose clustered organization was shown to be dependent on the spatial aggregation of an associated RNA binding protein, and necessary for asynchronous cell-cycle timing in multinucleate fungal cells [1].

Point processes can be defined as a special class of random measures, the *counting measures*, which assign non-negative integer values to all measurable subsets of a space. Spatial statistics such as the Ripley  $K$  function can be generalized to the framework of random measures; the  $K$  function generalizes directly

to the *reduced second moment measure* [6], which can be defined for stationary random measures taking either discrete values (counting measures) or continuous values. Further, the concept of spatial randomness can be generalized, leading to the class of *completely spatially random (CSR) measures*, which includes the homogenous Poisson processes as a subclass (those which are simultaneously counting measures and CSR). However, while such generalizations appear to enable the treatment of more general kinds of data, for instance continuous measurements which can be modeled as samples from a random measure, spatial statistics such as those above are rarely applied outside the point process context. Problems which arise in straightforwardly applying similar techniques to other kinds of data include choosing a general estimator for the  $K$  function, and determining a method to evaluate the necessary critical quantiles either by simulations or analytically for a general class of CSR null hypotheses. Unlike the homogeneous Poisson processes, which can be parameterized by a single intensity parameter, the class of all CSR random measures has a more complex structure, as characterized in [7,8]. In addition to the homogeneous Poisson processes, further subclasses of CSR random measures include Gamma processes, and sum measures associated with Marked Poisson processes (referred to as Mark Sum Poisson processes below).

We propose here a general approach to  $K$  function-based statistical tests in the context of arbitrary random measures. We provide a consistent convolution estimator for the  $K$  function based on the approach of [9], and investigate a number of ways in which the critical quantiles of the clustering index and degree of clustering estimators can be estimated for various classes of null model. First, we consider null hypotheses in the classes of Gamma processes and Mark Sum Poisson processes, and show how to fit these models to data and draw samples to simulate CSR in each case, providing an expectation-maximization (EM) algorithm to fit the Marked Poisson process. Further, we derive an exact permutation-based estimator for the clustering index, which provides a general test against the null hypothesis class of all CSR measures. We show that our permutation test using the convolution-based estimator reduces to the clustering index estimator used by Lee et. al. for the point process case [1], and hence provides a further rationale for the *conditionality principle* discussed in [4], which circumvents model fitting in the homogeneous Poisson case by fixing the number of points across simulations.

An advantage of adopting a general random measure based approach to identification of spatial patterning is that it provides a unifying framework in which statistics and indicators can be compared when analysing diverse data types. It also has the potential to provide a unifying framework for the modeling and inference of spatially distributed regulatory networks (at both inter- and intra-cellular levels) as diverse kinds of spatial omics data become available [10]. Random measures have emerged in a variety of areas of machine learning as a robust general framework for modeling diverse kinds of data, while avoiding the need to make arbitrary assumptions about the parametrization of distributions, particularly in context of Bayesian non-parametric approaches (see [8] for a general summary, and [11,12,13,14] for applications in text and image processing). We discuss in further detail below the potential relevance of our approach and the random measure framework within the broader context of modeling spatial omics data.

We begin by introducing formally the concepts of complete spatial randomness and random measures, and outline existing statistical tests for Ripley’s  $K$ ,  $L$  and  $H$  functions, the clustering index, and degree of clustering in the point process context (Sec. 2). We then outline our generalization of these tests to the context of arbitrary random measures, including a convolution-based estimator for the  $K$  function, and tests against various null hypothesis classes as described above (Sec. 3.1). We assess the ability of these tests to identify spatial randomness and patterning (clustering) first in synthetic data (Sec. 3.2), and then apply the method to probe for patterns of clustering over time in fluorescence microscopy data from pairs of corresponding mRNAs and proteins in a polarizing mouse fibroblast system (Sec. 3.3). The strong relationship between mRNA and protein clustering profiles suggests that mRNA localization and local translation provides a mechanism for protein localization in a number of cases, providing a small-scale demonstration of a spatial omics application. We conclude with a discussion (Sec. 4).

## 2 Preliminaries

### 2.1 Complete Spatial Randomness and Random Measures

A *random measure* can be defined on any measurable space  $\mathcal{S}$ , that is, a set equipped with a  $\sigma$ -algebra. For convenience, we will assume below that  $\mathcal{S}$  is a Euclidean space of dimension  $d$ , ( $\mathcal{S} = \mathbb{R}^d$ ), and that the  $\sigma$ -algebra is  $\mathcal{B}$ , the collection of Borel sets. A *Borel set* is any set that can be formed by the operations of countable union, countable intersection and relative complement from the open sets in the standard topology. A *measure* on  $\mathbb{R}^d$  is a mapping  $\phi$  from  $\mathcal{B}$  to the non-negative reals with infinity, such that  $\phi(\emptyset) = 0$ ,

and  $\phi(\cup_i B_i) = \sum_i \phi(B_i)$  for all countable collections of disjoint sets in  $\mathcal{B}$ ,  $\{B_i\}_{i \in \mathbb{N}}$ . A measure is called *locally finite* if  $\phi(B)$  is finite whenever  $B$  is a bounded set, and we denote the collection of all locally finite measures as  $\mathbb{M}$ . A *random measure* is then defined to be a random variable taking values in  $\mathbb{M}$ , and we will write  $\Phi$  for the random variable itself, and  $\phi$  for a specific value (measure) taken by  $\Phi$ . A random measure is necessarily defined with respect to a  $\sigma$ -algebra over  $\mathbb{M}$ , and all examples below will assume the  $\sigma$ -algebra  $\mathcal{M}$ , which is the smallest  $\sigma$ -algebra of subsets of  $\mathbb{M}$  such that all functions  $\phi \mapsto \phi(B)$  are measurable for arbitrary Borel set  $B$ . Further, we will use the notation  $P(\phi(B) \in R)$  to denote the probability that a random measure assigns a value in  $R$  to set  $B$ , where  $R$  is an open interval in  $\mathbb{R}$ .

A random measure is *completely random* if  $P(\phi(B_1) \in R_1)$  is independent of  $P(\phi(B_2) \in R_2)$  whenever  $B_1 \cap B_2 = \emptyset$ . *Complete Spatial Randomness* (CSR) is a stronger property of a random measure which implies both (a) complete randomness, and (b) *stationarity*,  $P(\phi(B) \in R) = (\phi(B+z) \in R)$  for any displacement  $z \in \mathbb{R}^d$ . A number of properties follow from complete spatial randomness. First, a CSR measure is necessarily isotropic, and there exists a fixed *intensity parameter*  $\lambda$  such that  $\mathbb{E}[P(\phi(B))]$  is  $\lambda \nu(B)$ , where  $\nu(B)$  is the *Lebesgue* measure on  $\mathbb{R}^d$ , which returns the volume of  $B$  [6], and  $\mathbb{E}[\cdot]$  denotes expectation. Further, any CSR measure over  $\mathbb{R}^d$  can be represented as a Poisson process  $\phi^*$  over  $\mathbb{R}^{d+1}$ , whose intensity measure has the form  $\lambda^*(B \times R) = \lambda_0 \nu(B) \gamma(R)$ , where  $\gamma(R)$  is a measure over  $\mathbb{R}$  (with  $\gamma(\mathbb{R})$  finite),  $\lambda_0$  is a non-negative real constant, and  $\phi(B) = \sum_{\mathbf{x} \in \phi^* \cap (B \times \mathbb{R})} x_{d+1}$  (see below for notational conventions for point processes). This follows from the general characterization of CSR measures given in [7] (see also [15,8]). A consequence of this representation is that  $P(\phi(B_1) \in R) = P(\phi(B_2) \in R)$  whenever  $B_1$  and  $B_2$  have equal volume,  $\nu(B_1) = \nu(B_2)$ , so that the distribution of  $\phi(B)$  is determined only by the volume of  $B$ .

A *point process* can be defined as a special type of random measure for which  $\phi(B) \in \mathbb{N} \cup \{0, \infty\}$  with probability 1, along with the technical condition that  $\phi(\{\mathbf{x}\}) \in \{0, 1\}$  for all  $\mathbf{x} \in \mathbb{R}^d$ , which ensures that no two points coincide (also called *simplicity*, [6]). Since point processes take only non-negative integer values on bounded subsets, they are also called *counting measures*. Further, since a sample  $\phi$  from a point process is (with probability 1) a countable subset of  $\mathbb{R}^d$  [6], we can use set notation and replace integrals by infinite sums in defining quantities for point processes, writing for example  $\phi(B) = \sum_{\mathbf{x} \in \phi \cap B} 1 = |\phi \cap B|$ . The class of CSR point processes is equivalent to the class of homogeneous Poisson processes. The homogeneous Poisson processes are parameterized by a single intensity parameter,  $\lambda$ , such that  $P(\phi(B) = n) = \text{Poisson}(n; \lambda \nu(B))$ , where  $\text{Poisson}(a; b) = (b^a / a!) \exp(-b)$  is the Poisson probability mass function. The more general class of Poisson processes (as used in the general characterization of CSR above) are completely random measures (without stationarity), parameterized by an *intensity measure*  $\lambda$  such that  $P(\phi(B) = n) = \text{Poisson}(n; \lambda(B))$  [6].

## 2.2 Statistical Tests for Spatial Patterning in Point Processes

For a stationary point process, the Ripley  $K$  function can be defined in terms of the *reduced second moment measure*  $\mathcal{K}$  [6,4]:

$$\begin{aligned} K(r) &= \mathcal{K}(B(o, r)) \\ \mathcal{K}(B) &= (1/\lambda) \mathbb{E}_{P_o}[\phi(B \setminus \{o\})], \end{aligned} \quad (1)$$

where  $o$  is the origin,  $B(o, r)$  is an open ball at the origin of radius  $r$ , and  $\mathbb{E}_{P_o}[\cdot]$  is the expectation under the *Palm distribution* at the origin, which for a stationary point process can be thought of as the original process conditioned on observing a point at  $o$ . Hence,  $P_o$  includes a point at  $o$  with probability 1, and this point is removed by taking the set difference  $\phi(B \setminus \{o\})$  when evaluating the reduced second moment measure.  $K(r)$  is therefore the expected number of further points observed within a radius  $r$  of an arbitrary point (due to stationarity) in  $\phi$ . Ripley's  $L$  and  $H$  functions can be defined in terms of  $K$  as:

$$\begin{aligned} L(r) &= \sqrt[d]{K(r) / \nu(B(o, r))} \\ H(r) &= L(r) - r. \end{aligned} \quad (2)$$

which have the effect of normalizing  $K$  so that for a homogeneous Poisson processes they take the form  $L(r) = r$  and  $H(r) = 0$ .

In [16], an estimator for  $K$  is proposed:

$$\hat{K}(r) = \frac{1}{\lambda^2 w(r)} \sum_{\substack{\mathbf{x}, \mathbf{y} \in \phi \cap W, \\ \mathbf{y} \neq \mathbf{x}}} [d(\mathbf{x}, \mathbf{y}) \leq r], \quad (3)$$

where  $[A]$  is the Iverson bracket, which is 1 when  $A$  is true and 0 otherwise,  $d(\cdot, \cdot)$  is the Euclidean distance,  $W \in \mathcal{B}$  is the window region in which the sample  $\phi$  is observed, and  $w(\cdot)$  is an *edge correction*:

$$w(r) = \mathbb{E}_{t_r}[\nu(W \cap (W + t_r))], \quad (4)$$

where  $t_r$  is a random vector sampled from a uniform distribution over the sphere centered at the origin of radius  $r$ . Eq. 3 is shown to be unbiased for all  $r$  less than the diameter of  $W$  for any convex  $W$  [16]. A simpler (but biased) estimator for  $K$  is also commonly used [6,4], which replaces the edge correction function with the volume/area of the observed region:

$$\tilde{K}(r) = \frac{1}{\lambda^2 \nu(W)} \sum_{\substack{\mathbf{x}, \mathbf{y} \in \phi \cap W, \\ \mathbf{y} \neq \mathbf{x}}} [d(\mathbf{x}, \mathbf{y}) \leq r]. \quad (5)$$

The associated statistical tests introduced below are unaffected by the choice between  $\hat{K}$  and  $\tilde{K}$ , and estimators for  $L$  and  $H$  can be straightforwardly derived from  $\hat{K}$  and  $\tilde{K}$  by substituting these estimators for true values in Eq. 2.

In [1], the *clustering index* statistic is introduced, which is denoted  $H^*$ . We provide a general expression for  $H^*$  below, which provides a test for clustering or dispersion at significance level  $\omega \in (0, 0.5)$ :

$$H^*(r) = \begin{cases} \frac{\hat{H}(r) - \hat{H}_{0.5}(r)}{\hat{H}_{(1-\omega)}(r) - \hat{H}_{0.5}(r)} & \text{if } (\hat{H}(r) \geq \hat{H}_{0.5}(r)) \wedge (\hat{H}_{(1-\omega)}(r) > \hat{H}_{0.5}(r)) \\ -\frac{\hat{H}_{0.5}(r) - \hat{H}(r)}{\hat{H}_{0.5}(r) - \hat{H}_\omega(r)} & \text{if } (\hat{H}(r) \leq \hat{H}_{0.5}(r)) \wedge (\hat{H}_{0.5}(r) > \hat{H}_\omega(r)) \\ 0 & \text{otherwise.} \end{cases} \quad (6)$$

where  $\hat{H}_\omega(r)$  denotes the  $\omega$ 'th quantile ( $(100\omega)$ 'th percentile) of  $\hat{H}(r)$  under an appropriate simulation of CSR (unlike [1], we use a median instead of a mean simulation-based estimator to center  $H^*(r)$ , so that  $H^*(r) = 0$  when  $\hat{H}(r) = \hat{H}_{0.5}(r)$ , to avoid complications arising if the mean estimator is greater than  $\hat{H}_{(1-\omega)}(r)$  or less than  $\hat{H}_\omega(r)$ ).  $H^*$  is thus a further normalization of  $\hat{H}$  such that, for a given value of  $r$ ,  $H^*(r) > 1$  iff  $H(r)$  (and hence  $K(r)$ ) is significantly above the range expected under CSR on a 1-sided test at level  $\omega$ , providing evidence of clustering (respectively,  $-H^*(r) > 1$  for dispersion) at length-scale  $r$ . By inspecting Eq. 6, we see that the edge correction terms from Eq. 3 will cancel in calculating  $H^*$  from  $\hat{K}$ , and hence it is sufficient to use the simpler estimator  $\tilde{K}$ .

To calculate  $\hat{H}_{(1-\omega)}(r)$  and  $\hat{H}_\omega(r)$  it is necessary to fix a distribution for simulations appropriate for the CSR null hypothesis. One possibility is to estimate the intensity parameter  $\lambda$  directly from  $\phi$  ( $\lambda = \phi(W)/\nu(W)$ ), and simulate a homogeneous Poisson process with this  $\lambda$  parameter by drawing first a Poisson distributed value  $N$  for the number of points in  $W$  from  $\text{Poisson}(N; \lambda\nu(W))$  for each simulation, and then distributing  $N$  points across  $W$  (independently and uniformly). This method is termed *parametric bootstrapping*, as discussed in [6,17,4], and provides an asymptotically consistent statistical test (as  $\nu(W) \rightarrow \infty$ ). Alternatively, we may condition all simulations on the number of points observed in  $\phi$ . Hence, we can take advantage of the *conditionality principle* discussed in [4], whereby the distribution of points in region  $W$  for any homogeneous Poisson process is independent of  $\lambda$  when conditioned on  $N$ . The points must be independently and uniformly distributed in  $W$  regardless of  $\lambda$ , forming a *binomial process* over  $W$  (see [6]). By conditioning on  $N$ , we therefore derive a consistent statistical test independent of the size of  $W$  against all CSR point processes (homogeneous Poisson processes), which is the approach taken in [6,4]. We note however that the simulations for the conditional test are no longer strictly CSR, since they are simulations of a binomial process. This distinction will be important in generalizing  $H^*$ . In particular, if an observation  $\phi$  is quantized across  $W$  into voxels which are small enough that the probability of two points occupying the same voxel is negligible, it is possible to view simulations of a binomial process as permutations of the voxels in  $W$ , and derive the binomial process test as a Monte-Carlo approximation to an exact permutation test, as will be proposed for the general case. The options discussed above for calculating  $H^*$  are summarized in Algorithm 1, which also serves as a template for generalization below (where  $\mathbf{X}$  denotes a spatially quantized observation of  $\phi$ ; here, a binary indicator vector across voxels lying in an observation window  $W$  which is 1 iff a voxel contains a point in  $\phi$ ). In [6],  $H^*$  is further used to define the *degree of clustering*  $\hat{\delta}(r) = \int_{t \in (0, r)} \max(H^*(t) - 1, 0) dt$ , which is the area of the curve  $H^*(\cdot)$  above 1 from 0 to  $r$ , and hence serves as an indicator for the degree of departure from CSR in this range.

---

**Algorithm 1** Generalized Estimator for Clustering Index,  $H^*$ 

---

**Input:**  $\mathbf{X}$  (Vectorized sample from point process / random measure),  $T$  (number of simulation / permutation trials),  $\omega$  (significance level)

- 1: Calculate estimators for  $K_X$ ,  $L_X$  and  $H_X$  using  $\mathbf{X}$  (Eqs. 2 and 5).
- 2: Draw vectorized samples  $\mathbf{Y}_1 \dots \mathbf{Y}_T$  using one of the following methods:
  - (a) (parametric bootstrapping) Find the best fitting CSR model  $M$  for  $\mathbf{X}$  in chosen null hypothesis class and run  $T$  simulations of  $M$ ,
  - (b) (conditioning) Simulate the null model  $T$  times conditioned on the measure or point count of the whole observed region in  $\mathbf{X}$ ,
  - (c) (permutation) Draw  $T$  permutations of  $\mathbf{X}$ .
- 3: Calculate  $K_{Y_t}$ ,  $L_{Y_t}$  and  $H_{Y_t}$  estimators on  $\mathbf{Y}_t$  for  $t = 1 \dots T$  (Eqs. 2 and 5).
- 4: Calculate  $\omega$ 'th,  $0.5$ 'th and  $(1 - \omega)$ 'th quantiles of  $H_{Y_1}(r) \dots H_{Y_T}(r)$  for each value of  $r$ , and use to normalize  $H_X$  to calculate the clustering index,  $H^*$  (Eq. 6).
- 5: **return**  $H^*$

---

### 3 Results

#### 3.1 Generalized Statistical Tests for Spatial Patterning in Arbitrary Random Measures

We now consider the generalization of the statistical tests and indices above from the point process case to the general random measure case. For a stationary random measure, the *reduced second moment measure*  $\mathcal{K}$  and Ripley's  $K$  function are defined exactly as in Eq. 1 (see [18], Eq. 2.19). The relevant *Palm distribution*,  $P_o$ , in the random measure case takes the form  $P_o(Y) = \int g(x) \mathbf{1}_Y(\phi + x) \phi(\mathrm{d}x) P(\mathrm{d}\phi)$ , with  $g(\cdot) : \mathbb{R}^d \rightarrow \mathbb{R}^+$  an arbitrary non-negative measurable function integrating to 1 and  $Y \in \mathcal{M}$  (see [9]). This definition can be seen to reduce to the distribution of further points conditioned on a point at the origin for the point process case, since  $g(x) \mathbf{1}_Y(\phi + x)$  will be non-zero only for  $x \in \phi$ , regardless of  $g(\cdot)$ . Ripley's  $L$  and  $H$  functions follow directly, as in Eq. 2.

To provide a general estimator for the Ripley  $K$  function, we must first specify how samples from the random measure  $\phi$  are observed. We assume that we have an observation window  $W$ , which can be partitioned into a collection of  $N$  regular cubical voxels with sides of length  $l$ , denoted  $v_1, v_2, \dots, v_N \subset W$ , whose centres lie at  $\mathcal{C} = \{c_1, c_2, \dots, c_N\}$ . Our observation of  $\phi$  is limited to the value it takes on each voxel, hence we observe the quantities  $\phi(v_1), \dots, \phi(v_N)$ . We can thus alternatively represent a sample as a measure  $\bar{\phi}$  with atoms at  $c_1, \dots, c_n$  having weights  $\phi(v_1), \dots, \phi(v_N)$  respectively. We now consider the estimator:

$$\begin{aligned} \bar{K}(r) &= \frac{1}{\lambda^2 \nu(W)} \sum_{n_1, n_2 \in \{1 \dots N\}} [\|c_{n_1} - c_{n_2}\| \leq r] \phi(v_{n_1}) \phi(v_{n_2}) - \bar{C} \\ &= \frac{1}{\lambda^2 \nu(W)} \int \int [\|x - y\| \leq r] \bar{\phi}(\mathrm{d}x) \bar{\phi}(\mathrm{d}y) - \bar{C} \\ \bar{C} &= \frac{\sum_{n=1 \dots N} (\bar{\phi}(v_n))^2}{\lambda^2 \nu(W)}. \end{aligned} \quad (7)$$

$\bar{K}(\cdot)$  can be efficiently calculated using a discrete convolution, since we have:

$$\int \int [\|x - y\| \leq r] \bar{\phi}(\mathrm{d}x) \bar{\phi}(\mathrm{d}y) = \int [\|x\| \leq r] (\bar{\phi} * \bar{\phi}')(x) \mathrm{d}x, \quad (8)$$

where  $\bar{\phi}'(\{x\}) = \bar{\phi}(\{-x\})$ , and  $(\bar{\phi} * \bar{\phi}')$  is the convolution of  $\bar{\phi}$  and  $\bar{\phi}'$  when treated as functions from  $\mathbb{R}^d$  to  $\mathbb{R}$ , hence  $\bar{\phi}(x) = \bar{\phi}(\{x\})$ .

$\bar{K}(\cdot)$  is an estimator for  $K(\cdot)$  in the following sense:

**Proposition 1.** For all values of  $r$ ,  $\bar{K}(\max(r - \sqrt{ld}, 0)) \leq \tilde{K}(r) \leq \bar{K}(r + \sqrt{ld}) + \bar{C}$ , where

$$\begin{aligned} \tilde{K}(r) &= \frac{1}{\lambda^2 \nu(W)} \int \int \mathbf{1}_W(x) \mathbf{1}_W(y) [\|x - y\| \leq r] \phi(\mathrm{d}x) \phi(\mathrm{d}y) - C \\ C &= \frac{\int_W \phi(\{x\}) \phi(\mathrm{d}x)}{\lambda^2 \nu(W)}. \end{aligned} \quad (9)$$

**Proof.** We begin by defining a function  $V : W \rightarrow \mathcal{C}$  such that we have  $x \in v_n$  implies  $V(x) = c_n$  (hence  $V$  sends  $x$  to the centre of the voxel to which it belongs). Then, we can rearrange Eq. 7 as follows:

$$\begin{aligned}\bar{K}(r) &= \frac{1}{\lambda^2 \nu(W)} \int \int [|x - y| \leq r] \bar{\phi}(\mathrm{d}x) \bar{\phi}(\mathrm{d}y) - \bar{C} \\ &= \frac{1}{\lambda^2 \nu(W)} \int \int \mathbf{1}_W(x) \mathbf{1}_W(y) [|V(x) - V(y)| \leq r] \phi(\mathrm{d}x) \phi(\mathrm{d}y) - \bar{C}.\end{aligned}\quad (10)$$

By inspection, the form of Eq. 10 is identical to Eq. 9 with the term  $[|V(x) - V(y)| \leq r]$  substituted for  $[|x - y| \leq r]$ , and  $\bar{C}$  substituted for  $C$ . Since each voxel is a  $d$  dimensional cube with sides of length  $l$ , we have  $\max(|x - V(x)|) = \sqrt{ld}/2$ . Hence, by the triangle inequality:

$$|V(x) - V(y)| - \sqrt{ld} \leq |x - y| \leq |V(x) - V(y)| + \sqrt{ld}.\quad (11)$$

Writing  $S(|x - y| \leq r)$  for the subset of  $W \times W$  for which  $[|x - y| \leq r] = 1$  (and similarly for  $S(|V(x) - V(y)| \leq r - \sqrt{ld})$  and  $S(|V(x) - V(y)| \leq r + \sqrt{ld})$ ), this implies:

$$S(|V(x) - V(y)| \leq r - \sqrt{ld}) \subseteq S(|x - y| \leq r) \subseteq S(|V(x) - V(y)| \leq r + \sqrt{ld}).\quad (12)$$

The three subsets in Eq. 12 correspond directly to the regions of integration for which the Iverson bracket conditions in Eqs. 10 and 9 are true ( $[.] = 1$ ) when the three quantities in the proposition are substituted (noting that  $r \geq 0$ ). The proposition follows from the nested relationship between these regions of integration, the fact that  $\phi$  is non-negative, and the fact that  $\bar{C} \geq C$  (since for any voxel  $v_n$ ,  $(\bar{\phi}(v_n))^2 = (\int_{v_n} \phi(\mathrm{d}x))^2 \geq \int_{v_n} \phi(\{x\}) \phi(\mathrm{d}x)$ ).

□

$\tilde{K}(r)$  is related to a further estimator  $\hat{K}$ , which substitutes  $(\mathbf{1}_W(x) \mathbf{1}_W(y)) / (w'(x - y))$  for  $\mathbf{1}_W(x) \mathbf{1}_W(y) / \nu(W)$  in Eq. 9, where  $w'(x - y) = \nu(W \cap (W + x - y))$  is an edge-correction term. We can derive  $\hat{K}$  as an unbiased estimator of  $K$  from fully observed (not spatially quantized) samples  $\phi$  using a result in [9] (using their Eq. 10, following Theorem 1, see Appendix A for the derivation). For values of  $r$  which are small compared to the diameter of  $W$ ,  $w'(x - y) \approx \nu(W)$  whenever  $[|x - y| \leq r] = 1$ , and hence  $\tilde{K}(r) \approx \hat{K}(r)$ ; hence  $\bar{K}(r)$  provides an approximation to  $\hat{K}(r)$  for small  $r$  up to the bounds in Prop. 1. We note also that  $\bar{K}(r)$  reduces (up to spatial quantization) to the previous estimator Eq. 5 in the point process case:

**Proposition 2.** For a point process sample  $\phi$ , for  $\bar{K}$  as in Eq. 7 we have:

$$\bar{K}(r) = \frac{1}{\lambda^2 \nu(W)} \sum_{\substack{\mathbf{x}, \mathbf{y} \in \bar{\phi} \cap W, \\ \mathbf{y} \neq \mathbf{x}}} [d(\mathbf{x}, \mathbf{y}) \leq r],\quad (13)$$

where  $x' \in \bar{\phi}$  iff  $x' = c_n$  for a voxel  $v_n$  for which there exists  $x \in \phi$  such that  $x \in v_n$ , and  $\lambda = \phi(W) / \nu(W)$  (where  $\phi(W)$  is the number of observed points). Additionally, we assume that the voxel width is chosen so that for no voxel  $\phi(v_n) > 1$ .

**Proof.** For the point process sample  $\phi$  as in the theorem (such that the voxel width is chosen so that for no voxel  $\phi(v_n) > 1$ ), we have:

$$\begin{aligned}\bar{K}(r) &= \frac{1}{\lambda^2 \nu(W)} \int \int [|x - y| \leq r] \bar{\phi}(\mathrm{d}x) \bar{\phi}(\mathrm{d}y) - \frac{\phi(W)}{\lambda^2 \nu(W)} \\ &= \frac{1}{\lambda^2 \nu(W)} \sum_{\substack{\mathbf{x}, \mathbf{y} \in \bar{\phi} \cap W \\ \mathbf{y} = \mathbf{x}}} [d(\mathbf{x}, \mathbf{y}) \leq r] - \frac{1}{\lambda},\end{aligned}\quad (14)$$

using the fact that  $\phi(v_n) \in \{0, 1\}$  implies  $\sum_{n=1 \dots N} (\bar{\phi}(v_n))^2 = \phi(W)$ , and  $\lambda = \phi(W) / \nu(W)$ . Further,

$$\begin{aligned}\frac{1}{\lambda^2 \nu(W)} \sum_{\substack{\mathbf{x}, \mathbf{y} \in \bar{\phi} \cap W, \\ \mathbf{y} = \mathbf{x}}} [d(\mathbf{x}, \mathbf{y}) \leq r] &= \frac{\phi(W)}{\lambda^2 \nu(W)} \\ &= \frac{1}{\lambda},\end{aligned}\quad (15)$$

and the proposition follows.

□

The estimator  $\bar{K}$  can be used in Algorithm 1 as above to calculate the  $H^*$  clustering index statistic for a general random measure, where  $\mathbf{X}$  here is identical to  $\bar{\phi}$ . All three options (parametric bootstrapping, conditioning and permutation) can be used in step 2 of the algorithm, and we consider below these options in connection with various null hypothesis classes. We begin by considering two restricted classes of CSR random measures as null hypotheses, stationary *Gamma* and *Mark Sum Poisson* processes, before considering options for the class of all CSR random measures.

**Gamma process.** A stationary *Gamma* process is defined as a random measure whose marginals are Gamma distributed as follows:

$$P(\phi(B)) = \text{Gamma}(\cdot; a\nu(B), b), \quad (16)$$

which is a CSR measure [8]. On the null hypothesis that  $\phi$  is a sample from a Gamma process,  $\phi(v_1), \dots, \phi(v_N)$  will be distributed according to  $\text{Gamma}(\cdot; a, b)$ , assuming for simplicity  $l = 1$  (the voxel sides are unit length) and hence  $\nu(v_n) = 1$  for all voxels. The parameters  $a$  and  $b$  can thus be set directly by fitting a Gamma distribution to  $\phi(v_1), \dots, \phi(v_N)$  by maximum likelihood [19], and the resulting Gamma process simulated by drawing independent identically distributed values from  $\text{Gamma}(\cdot; a, b)$  at each voxel. We note that it is also possible to simulate a Gamma process using a *stick-breaking* algorithm, as in [13], which may be more efficient if the number of voxels is large, with many taking values close to zero. Further, if it is assumed that  $b = 1$ , it is possible to choose the conditional version of step 2 in Algorithm 1, by first drawing  $N$  values from  $\text{Gamma}(\cdot; a, 1)$ , and normalizing to sum to  $\phi(W)$ , the observed sample total. This is equivalent to simulating a *Dirichlet process* with intensity parameter  $a$ , and scaling by  $\phi(W)$ , for which it is possible also to use a stick-breaking algorithm [8].

**Mark Sum Poisson process.** A *Mark Sum Poisson* process can be defined as a random measure whose marginals are distributed as:

$$P(\phi(B) \in R) = \sum_{\mathbf{n} \in (\mathbb{N} \cup \{0\})^M} [(\sum_m w_m n_m) \in R] \cdot \prod_m \text{Poisson}(n_m | \alpha_m \nu(B)), \quad (17)$$

where  $m = 1 \dots M$  are the *marks* of the process, each associated with a weight  $w_m \geq 0$  and intensity  $\alpha_m > 0$ , and  $[A]$  is the Iverson bracket, which is 1 when  $A$  is true and 0 otherwise. The process is so-called, since it is equivalent to attaching marks to the points in a homogeneous Poisson process with intensity  $\lambda = \sum_m \alpha_m$ , where mark  $m$  appears with a probability proportional to  $\alpha_m$  (forming a *Marked Poisson process*), and the value  $\phi(B)$  is calculated by summing across the weights  $w_m$  of the points in  $B$  (forming its associated *sum measure*) [6]. In this equivalent representation, each mark independently follows a homogeneous Poisson processes with intensity  $\alpha_m$ , and hence it follows that Eq. 17 is CSR [6]. On the null hypothesis that  $\phi$  is a sample from a Mark Sum Poisson process,  $\phi(v_1), \dots, \phi(v_N)$  will be distributed according to Eq. 17, assuming  $l = 1$ , which we call a *Weighted Sum of Poisson distributions*. By fixing the weights  $w_1 \dots w_M$ , it is possible to derive an expectation-maximization (EM) algorithm to fit  $\alpha_1 \dots \alpha_M$  by maximum-likelihood (see Appendix B). Having fitted the model, CSR samples can be drawn by generating values  $p_1, \dots, p_M$ , distributed as  $\text{Poisson}(\cdot | \alpha_1) \dots \text{Poisson}(\cdot | \alpha_M)$  respectively, and calculating  $\sum_m p_m w_m$  at each voxel.

Aside from forming a broad CSR measure class, Mark Sum Poisson processes are interesting in that, in the limit of infinite marks, they form a universal representation for CSR measures. This is because, as noted earlier, any CSR measure  $\phi$  over  $R^d$  can be represented as a (non-homogeneous) Poisson process  $\phi^*$  over  $R^{d+1}$  with intensity measure  $\lambda^*(B \times R) = \lambda_0 \nu(B) \gamma(R)$  such that  $\phi(B) = \sum_{\mathbf{x} \in \phi^* \cap (B \times \mathbb{R})} x_{d+1}$  [7]. As the number of marks increases, the  $\alpha_m$ 's are better able to approximate the measure  $\gamma$ , and hence any random measure. Although we considered above only the case of fitting a distribution with finite marks and fixed weights, by using a large number marks with densely and evenly sampled weights, it is therefore possible to approximate any CSR measure.

**General case.** We know that, since all voxels have identical volume  $\nu(v_n) = l^d$ ,  $\phi(v_1), \dots, \phi(v_N)$  are independent samples from the same distribution ( $P(\phi(B_1) \in R) = P(\phi(B_2) \in R)$  whenever  $\nu(B_1) = \nu(B_2)$ ). Hence, we can use the empirical distribution of voxel values as an estimate for  $P(\phi(B))$ ,  $\nu(B) = l^d$ , which is completely general in that the only assumption we have made is that  $\phi$  is CSR. We can thus approximate a simulation of CSR from the ‘best fitting’ CSR measure (whose marginals approach  $\phi$  asymptotically), by generating values for new voxels in the simulation using sampling with replacement of the values

$\phi(v_1), \dots, \phi(v_N)$  already seen (equivalently, sampling from the empirical distribution). We note that this only approximates CSR, since, with probability 1,  $\phi(v_n)$  takes a value in the empirical distribution, and hence the values taken by  $\phi$  on any disjoint set of sub-voxels which cover a given voxel must be dependent.

If instead of sampling with replacement from the empirical distribution to generate new samples, we permute the voxel values  $\phi(v_1), \dots, \phi(v_N)$  (sampling without replacement),  $\phi(W)$  must remain unchanged, and we can regard this as approximate sampling from the best fitting CSR measure conditioned on  $\phi(W)$ . However, rather than viewing permutation as an approximate simulation of CSR, it is also possible to view it in terms of an exact test against the general CSR null hypothesis, based on the exchangeability of the voxels under CSR. We summarize this as:

**Proposition 3.** *Algorithm 1 with exhaustive permutation at step 2, is an exact test for CSR of an arbitrary random measure at significance level  $\omega$ , in the sense that  $P(H^*(r) > 1) < \omega$  for an arbitrary distribution over the class of all CSR measures.*

**Proof.** Given random measure  $\phi$  over  $\mathbb{R}^d$ , observation window  $W$  and  $N$  cubical voxels with sides of length  $l$  partitioning  $W$ ,  $v_1, v_2, \dots, v_N$ , we can construct a related random measure  $\phi'$  over  $\mathbb{R}^N$  such that:

$$\phi'(B') = P([\phi(v_1), \phi(v_2), \dots, \phi(v_N)] \in B'), \quad (18)$$

where  $[a_1, a_2, \dots, a_N]$  denotes a vector in  $\mathbb{R}^N$ . Hence,  $\phi'(B')$  is the probability that  $\phi$  gives a combination of values to voxels  $1 \dots N$  lying in  $B'$ , where  $B'$  is a Borel set over  $\mathbb{R}^N$ . Further, we introduce the *rejection function*,  $f : \mathbb{R}^N \rightarrow \{0, 1\}$ , which takes the value 1 when  $H^*(r) > 1$  (for a fixed  $r$ ) using an exhaustive permutation test at step 2 of Algorithm 1 and significance level  $\omega$  at step 4, and 0 otherwise. Then, for any random measure:

$$P(H^*(r) > 1) = \int f(\mathbf{x}) \phi'(\mathrm{d}\mathbf{x}). \quad (19)$$

Considering now a CSR measure, by exchangeability of voxel regions, for any  $B'$  we have  $\phi'(B) = \phi'(\pi(B'))$  for all  $\pi \in \mathbb{P}$ ; where  $\mathbb{P}$  is the set of all permutations on  $N$  elements, and we let  $\pi([x_1, x_2, \dots, x_N]) = [x_{\pi(1)}, x_{\pi(2)}, \dots, x_{\pi(N)}]$  and  $\pi(B) = \{\mathbf{y} | \exists \mathbf{x} \in B \text{ s.t. } \pi(\mathbf{x}) = \mathbf{y}\}$ . Hence, now considering the region  $R = \{\mathbf{x} \in \mathbb{R}^N | x_1 \leq x_2 \leq \dots \leq x_N\}$ , under the assumption that  $\phi$  is CSR, we can rewrite Eq. 19 as:

$$\begin{aligned} P(H^*(r) > 1) &= \int [\mathbf{x} \in R] f^*(\mathbf{x}) \phi'(\mathrm{d}\mathbf{x}) \\ &= \int [\mathbf{x} \in R] \frac{f^*(\mathbf{x})}{g(\mathbf{x})} \phi''(\mathrm{d}\mathbf{x}), \end{aligned} \quad (20)$$

where,

$$\begin{aligned} f^*(\mathbf{x}) &= \sum_{\mathbf{y} \in \mathbb{P}(\mathbf{x})} f(\mathbf{y}) \\ g(\mathbf{x}) &= |\mathbb{P}(\mathbf{x})| \\ \phi''(B') &= \int g(\mathbf{x}) \phi'(\mathrm{d}\mathbf{x}), \end{aligned} \quad (21)$$

and we write  $\mathbb{P}(\mathbf{x})$  for the set  $\{\mathbf{y} | \exists \pi \in \mathbb{P} \text{ s.t. } \pi(\mathbf{x}) = \mathbf{y}\}$ . By definition of  $f$  we have that for all  $\mathbf{x}$ :

$$\frac{f^*(\mathbf{x})}{g(\mathbf{x})} < \omega. \quad (22)$$

Hence (using  $\phi''(R) = 1$ , by definition of  $g$ ):

$$\begin{aligned} P(H^*(r) > 1) &= \int [\mathbf{x} \in R] \frac{f^*(\mathbf{x})}{g(\mathbf{x})} \phi''(\mathrm{d}\mathbf{x}) \\ &< \int [\mathbf{x} \in R] (\omega) \phi''(\mathrm{d}\mathbf{x}) \\ &= \omega. \end{aligned} \quad (23)$$



□

As mentioned, Prop. 3 sheds further light on the practice of fixing the number of points  $N$  during point process simulations as in [4]. Further, we note that in practice, Monte Carlo sampling is typically required in place of exhaustive permutation in evaluating  $H^*$  using Algorithm 1. The main advantage of Prop. 3 is that it sidesteps the issues of choosing a particular CSR measure or null hypothesis class, and provides a justification for methods which do not simulate CSR exactly in the general case.

### 3.2 Results on Synthetic Data: Gamma process, Mark Sum Poisson process, and Poisson process with Gaussian Kernel

We first test our approach on synthetically generated data, where we are interested in determining if the various forms of Algorithm 1 can distinguish between data which is known to be completely spatially random, and data which is known to contain spatial structure in the form of clustering. For synthetic CSR data, we consider the Gamma process and Mark Sum Poisson process as discussed above, where the techniques used to draw samples from these processes discussed in the context of Algorithm 1 can be likewise be used to generate data for a synthetic test set. We sample  $a$  and  $b$  uniformly in the intervals  $(0, 10]$  and  $(0, 2]$  respectively for the Gamma process (see Eq. 16), and use five marks with the fixed weights 0.25, 0.5, 1, 2, 4 for the Mark Sum Poisson process (which for convenience we also fix during testing) while sampling  $\alpha_m$ 's uniformly in the interval  $[0.37, 2.7]$  (see Eq. 17). We also generate CSR data from a Poisson process with  $\lambda = 0.1$  (equivalently, a Mark Sum Poisson process with  $M = 1, \alpha = 0.1, w = 1$ ).

For a simple synthetic model with spatial structure, we use a model we describe as a *Poisson process with Gaussian Kernels*. This model has two parameters:  $\lambda$ , the intensity of an underlying Poisson process  $\phi_{\text{Poisson}}$ , and  $\sigma$ , the standard deviation of a Gaussian kernel attached to each point in  $\phi_{\text{Poisson}}$ . As a random measure, the model can be defined as follows:

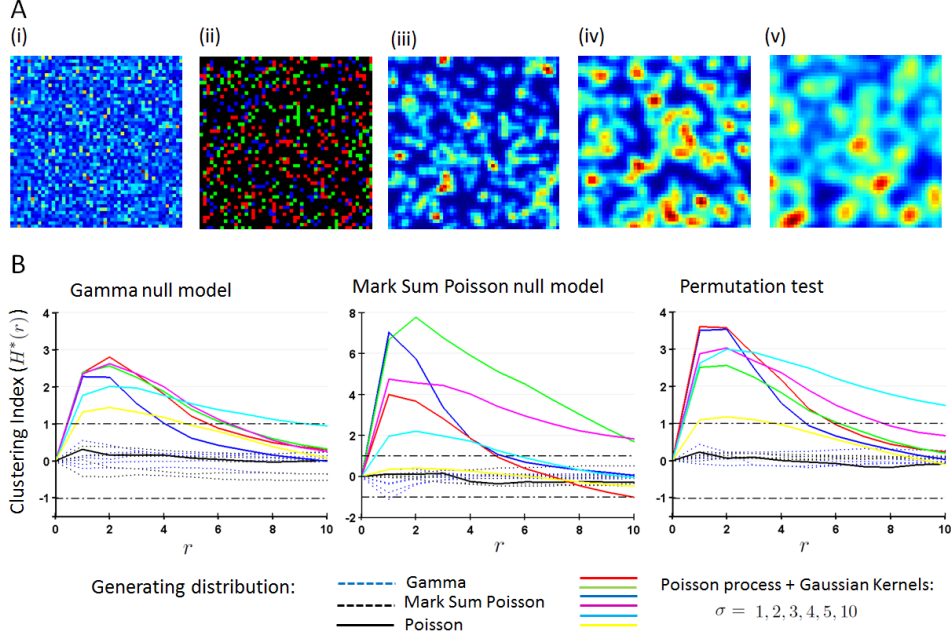
$$\phi(B) = \sum_{\mathbf{x} \in \phi_{\text{Poisson}}} \int_B \mathcal{N}(\mathbf{y}; \mathbf{x}, \sigma^2) d\mathbf{y}, \quad (24)$$

where  $\mathcal{N}(\cdot; \mu, \sigma^2)$  is the Gaussian probability density function with mean  $\mu$  and diagonal covariance  $\sigma I_d$  ( $I_d$  being the  $d$  dimensional identity matrix). Samples are drawn from this process by first drawing a sample from a Poisson process with intensity  $\lambda$  in a large region surrounding the simulation region, and for the  $n$ 'th voxel in the observed area summing across the values  $\mathcal{N}(c_n; \mathbf{x}, \sigma^2)$ , where  $c_n$  is the centre of voxel  $n$ , and  $\mathbf{x}$  ranges across the points from the Poisson process in the enclosing region. We fix  $\lambda = 0.1$ , and let  $\sigma$  take the values 1, 2, 3, 4, 5, 10 to simulate clustering at various length-scales.

All processes are simulated on a 2D window of  $50 \times 50$  pixels, with five examples generated under each setting. Examples from each process are shown in Fig. 1A. Algorithm 1 is run in three variations on all synthetic samples, the first two using the restricted Gamma and Mark Sum Poisson process classes respectively to fit and test against in step 2 (option (a)), and the third using the general permutation test (option (c)). Ripley  $K$  and clustering index functions are calculated for  $r = 0 \dots 10$  pixels, hence restricting  $r$  to relatively small values under which edge correction effects are expected to be minimized, as discussed above. In fitting the Mark Sum Poisson process, 5 iterations of the EM algorithm outlined in the Methods section are used, and 20 simulations/permutations are used to estimate the 5-95th percentile range of  $H(r)$  under CSR for each test sample and all algorithm variations ( $\omega = 0.05$ ).

Results on synthetic data as described are shown in Fig. 1B. In general, all three versions of the algorithm are able to discriminate correctly between CSR and clustered data. For the Gamma and Mark Sum Poisson process restricted null models, as expected, when the same class of models is used for testing and as null hypothesis, the clustering index stays well within the CSR region (Fig. 1B, left and central graphs). However, the results also show these models to be robust as CSR null hypotheses, and in each case test data from the other model and the simple Poisson process also generally stay within the CSR region. Samples from the Gamma process are marginally more dispersed when tested against the Mark Sum Poisson null model (Fig. 1B, central graph), which may be due to limitations in fitting the Mark Sum Poisson model with only five marks and the EM algorithm which only achieves a local optimum. The permutation test version of the algorithm appears to be the most robust, with all CSR test samples generating clustering index functions tightly located around zero (Fig. 1B, right graph).

With respect to the clustered data generated from the Poisson process with Gaussian Kernels, all three algorithmic variations are generally able to detect that the data is not CSR, with the clustering index breaching the +1 line (the median clustering index is shown for the five replications of each process, Fig. 1B



**Fig. 1. Detecting spatial patterning in synthetic data.** (A) Simulations of various random processes. (i) Sample from a Gamma process (scale is from blue to red for low to high values); (ii) sample from a marked Poisson process (red/green/blue represent marks with different weights); (iii-v) samples from a Poisson process with Gaussian kernels, where the kernel parameter  $\sigma = 1, 2, 4$  for images (iii)-(v) respectively (scale as in (i)). (i-ii) are CSR processes, while (iii-v) exhibit clustering at increasing length-scales. (B) Clustering index calculated on synthetic data from processes in (A). Horizontal lines at -1 and 1 show 5th and 95th percentiles of CSR simulations after normalization. These are derived from Gamma process simulations (left), Mark sum Poisson process simulations (centre), and permutations of the test sample (right).

all graphs). The only exception is the Mark Sum Poisson null model, where the clustering index for the  $\sigma = 10$  process remains within the CSR region, possibly indicating again difficulty in fitting the marginal distributions sufficiently to distinguish the clustering over large distances from CSR (notably the  $\sigma = 5$  line is also close to CSR for the Mark Sum Poisson null model, Fig. 1B central graph). Both Gamma process and permutation based variations detect clustering at all  $\sigma$  values. However, we note that the permutation test variation achieves a better separation of length-scales associated with the different  $\sigma$  values: generally the clustering index peaks earlier and returns to the CSR region earlier for the lower the  $\sigma$  value is, while all functions have similar profiles in the Gamma process case, except for  $\sigma = 5$  and 10 (Fig. 1B left and right graphs; similarly, the separation pattern is not strongly observed for the Mark Sum Poisson variation, central graph). This could be due to difficulties in fitting close enough Gamma distributions to the marginals of these processes to distinguish fine differences in spatial patterning. In general then, the permutation test variation of Algorithm 1 is shown to be robust in its ability to distinguish CSR from spatial patterning over various length-scales, and the variations using restricted classes of CSR null hypotheses are not shown to offer substantial advantages even in the case that the test data is from the matching CSR class for the data considered (while the substantially similar performance of all algorithmic variations is perhaps surprising given the restrictions on the null model imposed by the Gamma and Mark Sum Poisson process variations).

### 3.3 Results on Fluorescent Microscopy Data: Identifying Patterns of mRNA and Protein Localization over Time in a Polarizing Mouse Fibroblast System

We further tested our approach by using it to probe for spatial patterning in the subcellular distributions of mRNAs and proteins in a polarizing mouse fibroblast system. We used high resolution confocal microscopy to generate 3D data specifying individual mRNA positions using Fluorescence In Situ Hybridization (FISH) followed by spot detection, and protein abundance across at a grid of voxel positions using Immunofluorescence (IF). Such data thus allows us to test the ability of our algorithm to detect spatial patterning in both

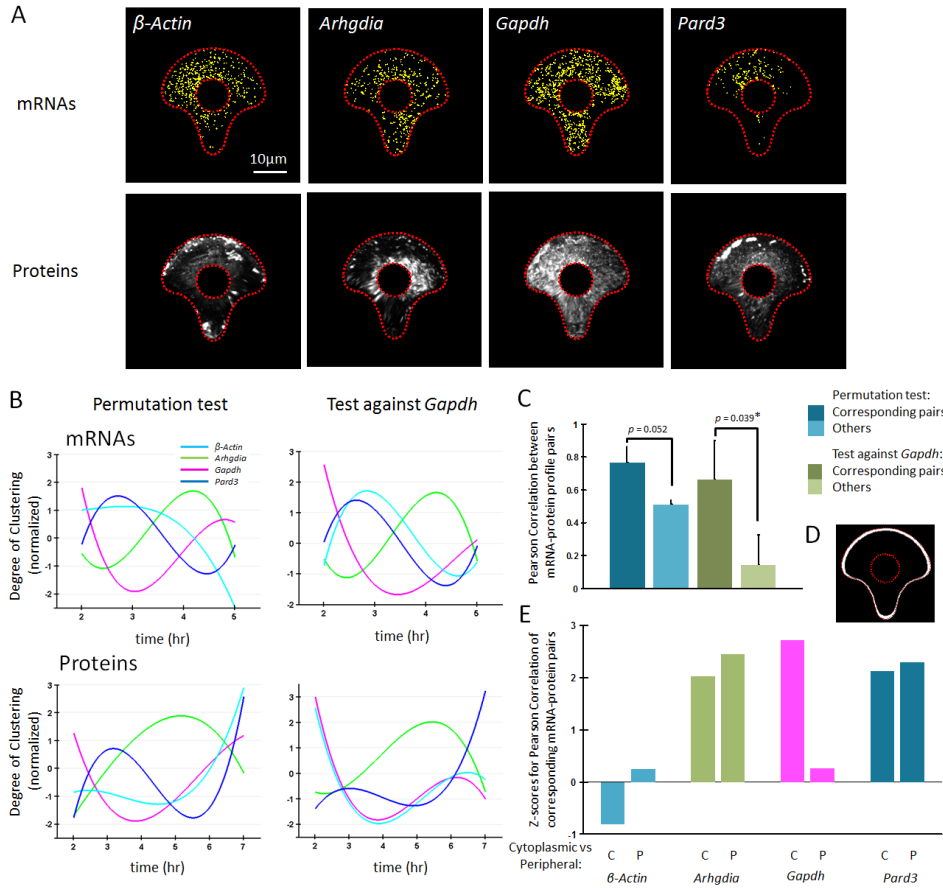
point data (mRNAs) and continuous valued data (protein intensities), which can be modeled similarly as random measures.

Cells were grown on cross-bow shaped micropatterns in order to standardize cell morphology and internal organelle arrangement (see [21,20]). The cells were serum-starved for 16 hours prior to micropatterning. Micropatterns were plated with Fibronectin, which causes the cells to begin to polarize following adhesion. Cells were then fixed in formaldehyde at various times post adhesion, and FISH or IF probes introduced to generate data for the distributions of four mRNAs, *Arhgdia*, *Pard3*,  $\beta$ -*Actin* and *Gapdh* and corresponding protein products. *Arhgdia* and *Pard3* were chosen, since they have been shown previously to exhibit significant spatial patterning in fibroblasts at the mRNA and protein [22] and protein only levels [23] respectively;  $\beta$ -*Actin* is known to exhibit spatial patterning at mRNA and protein levels in a variety of cell types [24]; and the house-keeping gene *Gapdh* is not expected to exhibit strong spatial patterning. We were particularly interested in the case of *Pard3* to investigate whether spatial patterning could be detected at the mRNA level, as had been shown the protein level previously. FISH data for mRNAs was collected 2, 3, 4 and 5 hours after micropatterning, and IF protein data at 2, 3, 5 and 7 hours, with  $\sim 40$  cells imaged per mRNA/protein at each of these time-points. Individual micropatterned cells were imaged as separate  $z$ -stacks ( $512 \times 512$  pixels, 15-25  $z$ -levels, with approximately  $0.1 \mu\text{m}$  pixel width and  $0.3 \mu\text{m}$  separation between  $z$ -levels). In addition, IF Tubulin staining was applied to all cells to identify the microtubule cytoskeleton, which enabled a simple cell-volume model to be constructed by identifying a 2D cell boundary and height map, and DAPI staining was applied to identify the nuclear region. Examples of data from each mRNA and protein are shown in Fig. 2A, which have been 2D projected and warped to an average micropattern shape for visualization. Further details on the experimental protocol are provided in Appendix C.

We applied Algorithm 1 to the 3D mRNA point data, and the protein IF to calculated clustering index functions and the degree of clustering (the area of the clustering index graph above +1, indicating significant clustering) from each cell individually. We applied two forms of the algorithm: in the first, we used the 3D Ripley  $K$  estimator, and applied the permutation method at step 2 to generate simulation samples, where permutations are applied to the voxels falling within the cell-volume model; in the second, we projected the points/summed intensities into 2D and applied the Ripley  $K$  estimator in 2D, while using the empirical distribution of *Gapdh* mRNAs and proteins to simulate CSR in step 2 (independently sampling, at each pixel, a binary value or intensity for mRNAs and proteins respectively from the same pixel in a *Gapdh* distribution, and normalizing the resulting sample to sum to the same total as the input sample; this is an empirical variation on option (b) of step 2). 100 simulations/permutations are used to estimate the 5-95th percentile range of  $H(r)$  under CSR for each test sample and both algorithm variations ( $\omega = 0.05$ ).

The profiles for the degree of clustering across time are shown for mRNAs and proteins using both versions of the algorithm in Fig. 2B. Cubic splines are fitted to the profiles for ease of visualization. Similarity in the profiles from the two algorithmic variations are readily apparent. Also visually apparent is a similarity between several of the mRNA profiles and the profiles of their protein products, particularly *Arhgdia*, *Gapdh* and *Pard3*. To further investigate the relationship between mRNA and protein profiles, we calculated the Pearson Correlation coefficient between pairs of mRNA-protein degree of clustering profiles for each algorithm version (matching mRNA time-points 2, 3, 4 and 5 to protein time-points 2, 3, 5 and 7 respectively), and tested for whether the correlations between mRNAs and their corresponding proteins were significantly higher than between randomly matched pairs. We found this was the case for both algorithm variations, although the difference was more pronounced for the second variant ( $p = 0.039$  versus  $p = 0.052$ , see Fig. 2C). The above suggests that our approach is able to identify significant aspects of spatial patterning in this system. Particularly, since the proteins are observed over a longer time period than the mRNAs, the similarities in profile reflect a stretching of this profile in the proteins with respect to the mRNAs. Plausibly, the spatial patterning at the mRNA level acts as a determinant of the patterning at the protein level through processes such as local translation, although this cannot be established directly from our approach. We note also that the more significant relationships observed for the second algorithmic variation may reflect the difficulties in estimating an accurate 3D cell volume model in the first version, which is required to select the voxels to be permuted in the simulations.

To probe the spatial patterning of the mRNA-protein pairs further, we compared the clustering profiles of each corresponding pair individually. We were interested also in gaining information about where in the cell clustering was occurring for each pair. To this end, in addition to the correlations between clustering profiles across the whole cytoplasm as above, we calculated also the correlations between clustering profiles using the mRNA and protein data restricted to a small peripheral region of the cytoplasm (see Fig. 2D), formed from a strip around the boundary of the cell which was 10% of the radial distance to the nucleus centroid in width (projected across all  $z$ -slices), hence reflecting peripheral clustering only. We chose this region, since



**Fig. 2. Detecting spatial patterning in FISH and IF data from polarizing micropatterned mouse fibroblast cells.** (A) Left column shows representative mRNA detections from FISH probes targeting transcripts of four genes. Right column shows representative IF distributions for protein products of these transcripts (IF intensity corresponding to protein abundance). Dotted lines show the average micropatterned cell shape and nucleus boundary, and cytoplasmic transcript locations and IF intensities are warped to this average shape for visualization. (B) The degree of clustering statistic is calculated for four mRNAs (at 2, 3, 4 and 5 hour time-points), and their protein products (at 2, 3, 5 and 7 hour time-points). CSR is simulated using a permutation test (left column) or the empirical distribution of *Gapdh* (right column). Median values are calculated at each time-point, and cubic splines are fitted after subtracting the mean and normalizing by the standard deviation across time for ease of visualization (each mRNA/protein normalized independently). (C) The Pearson correlation coefficient is calculated between all mRNA and protein pairings (including corresponding and non-corresponding pairs) based on the median degree of clustering profiles from (B), matching mRNA 2, 3, 4 and 5 hour time-points to protein 2, 3, 5 and 7 hour time-points respectively. The *Gapdh*-based CSR test leads to a significant separation of corresponding pairs (involving an mRNA and its protein product) versus others (one-tailed Mann-Whitney test). (D) Peripheral region, as described in text, shown in white. (E) Z-scores are calculated for the Pearson Correlation coefficient between the clustering profiles of each pair of corresponding mRNAs and proteins (using the permutation test based degree of clustering). The Z-score is calculated by comparing the correlation of a given corresponding pair to the distribution of correlations for all other pairs containing one member (mRNAs or proteins) of the given pair. Shown are the Z-scores calculated using the degree of clustering values for cytoplasmic and peripheral only populations.

*Arhgdia*, *Pard3*, and  $\beta$ -Actin proteins are known to localize peripherally (as is visually apparent in Fig. 2A), while localized translation at the periphery is known to occur in the cases of *Arhgdia* and  $\beta$ -Actin [22,24]. In both cases, we use the permutation test version of the algorithm.

Fig. 2E compares the individual profile correlations in the cytoplasmic and peripheral populations for each corresponding mRNA-protein pair, by Z-scoring the correlation of each corresponding pair against the correlations of non-corresponding mRNA-protein pairings (see figure legend). We observe that the *Gapdh* correlations are only strong when calculated across the whole cytoplasm, but disappear at the periphery. Although we did not expect strong spatial patterning in either case, the correlations observed in the cytoplasm may be due to the *Gapdh* mRNAs and proteins forming diffuse large-scale regions of higher concentration

in certain directions in response to polarization, which is suggested visually in the protein case (Fig. 2A; hence using *Gapdh* as a model of spatial randomness as above can be expected to mask clustering at such scales). In contrast, for both *Arhgdia* and *Pard3* we observe correlated mRNA-protein clustering dynamics in both the cytoplasmic and peripheral populations. This is expected in the case of *Arhgdia*; however, in the case of *Pard3*, clustering at the periphery has previously been demonstrated only at the protein level in a fibroblast system [23]. Our results suggest that peripheral *Pard3* mRNA localization is also important in this system, although as above we cannot directly conclude from our data that local translation establishes the protein localization (since the peripheral mRNA and protein clusters could form independently). Unexpectedly,  $\beta$ -Actin does not show significant correlation in the cytoplasmic or peripheral populations. We suggest that, while peripheral local translation may be occurring leading to peripheral protein clustering (as can be seen in Fig. 2A), since the  $\beta$ -Actin mRNA is highly dispersed we do not observe direct correlations between mRNA and protein clustering (the Z-score in Fig. 2E is high only if clustering occurs in both mRNAs and proteins and is correlated over time, and does not indicate the strength of independent clustering in mRNAs or proteins).

## 4 Discussion

We have described and analysed a general algorithm for calculating the Ripley- $K$  derived clustering index and degree of clustering statistics in the context of arbitrary random measures. Our approach generalizes the point-process-based approach in [1], while shedding further light on this approach and the conditionality principle noted in [4] by analysing these statistics as permutation tests in the random measure context, using the exchangeability of elements with identical volume in the context of completely spatially random measures. Through studies on synthetic data, we compared variations of the algorithm which explicitly simulate CSR using Gamma process and Mark Sum Poisson processes against the permutation approach, and found all variations were able to discriminate CSR from spatial patterning (clustering) at various length-scales in the data used, with the permutation-based algorithm offering a marginally more robust approach. Tests on fluorescence microscopy data from polarizing fibroblasts showed that the random-measure-based approach was able to identify spatial patterns in subcellular mRNA and protein distributions which were significantly correlated for corresponding mRNAs and protein products, hence suggesting that the patterns uncovered are biologically significant in the system and possibly reflective of mRNA/protein localization mechanisms tied to local translation.

We note that the random measure framework provides a useful theoretical context in which to frame problems in the spatial domain which require integration of diverse data-types, as is necessary for instance in spatially resolved omics problems [10]. Although we specifically analysed statistics based on the Ripley- $K$  function (providing a convolution-based estimator appropriate for the random measure context, Eq. 7), similar permutation-based tests can be applied to arbitrary statistical estimators to detect other kinds of patterning in a random measure context, using the exchangeability properties of CSR measures [8]. A possible use of such tests is in identifying functionally related genes from their spatial patterning in a given system, as for instance in the studies of [25] and [26]. The application of our approach to a polarizing fibroblast system above provides an example of how random measures may be used in such a context. We have concentrated here on deriving statistics from individual distributions which may be correlated to identify functional relationships; however the problem of identifying dependencies between spatial patterning across multiple distributions may be framed more generally in a random measure context, using for instance cross-correlation measures [9]. A further possible application of random measures and related models is in the inference of spatially distributed regulatory networks. Gaussian process (GP) models have already been applied in the context of modeling transcription factor networks, using a theoretical framework which allows models based on deterministic differential equations to be embedded in the GP covariance function for probabilistic inference [27], and are possibly also a suitable model for spatially distributed networks (we note though that GPs are not strictly random measures, since a function drawn from a GP may take negative values with non-zero probability, and hence cannot be treated directly as a density function). Alternatively, dynamic Dirichlet and Gamma process models [13] potentially offer other attractive ways of formulating the general spatial network inference problem. Within this context, random measure based statistics such as those investigated in this study may serve as evidence for or against causal relationships when combined with perturbations (as in [28]) or single-cell level multiplexed data (as in [29]). The above problems may therefore benefit directly from the techniques and analysis outlined the present study, as well as offering a broad set of challenges for future work drawing on similar theoretical approaches.

## Appendix A. Derivation of $\hat{K}$ from Stoyan and Ohser's Estimator for the Cross Correlation Measure

Stoyan and Ohser [9] consider the case of a stationary *weighted random measure*. A special case of this is a random measure over the space  $\mathbb{R}^d \times \mathbb{R}$  for which  $P(\psi(B \times U)) = P(\psi((B + t) \times U))$  for any  $B \in \mathcal{B}$  and  $U \in \mathcal{U}$  (with  $\mathcal{U}$  any  $\sigma$ -algebra over  $\mathbb{R}$ ). In their Eq. 4, they introduce the *reduced correlation measure*,  $\mathcal{K}_{12}$ , which can be used to express the second moment measure of  $\phi$ :

$$\mu^{(2)}(B_1 \times U_1 \times B_2 \times U_2) = \lambda(U_1)\lambda(U_2) \int_{B_1} \int \mathbf{1}_{B_2}(x + h) \mathcal{K}_{12}(dh) dx, \quad (25)$$

where  $\lambda(U_1)$  is the intensity of the stationary random measure  $\phi_{U_1}(B) = \phi(B \times U_1)$ . Stoyan and Ohser provide the following estimator for  $\mathcal{K}_{12}$  in their Eq. 10, which, following their Theorem 1, is shown to be unbiased:

$$\hat{\mathcal{K}}_{12}(B) = \frac{1}{\lambda(U_1)\lambda(U_2)} \int \int \frac{\mathbf{1}_B(y - x) \mathbf{1}_{W_1}(x) \mathbf{1}_{W_2}(y)}{\nu(W_1 \cap W_2 + x - y)} \mathbf{1}_{U_1}(w_1) \mathbf{1}_{U_2}(w_2) \phi(dx, w_1) \phi(dy, w_2). \quad (26)$$

Any random measure over  $\mathbb{R}^d$  can be considered a weighted random measure where  $\mathcal{U}$  is taken to be the trivial  $\sigma$ -algebra,  $\mathcal{U} = \{\emptyset, \mathbb{R}\}$ . Hence, considering the case that  $W_1 = W_2 = W$  and writing  $\phi(B)$  for  $\phi(B \times \mathbb{R})$ ,  $\lambda$  for  $\lambda(\mathbb{R})$ , and  $\mathcal{K}_{11}$  for what we shall call the *reduced autocorrelation measure* (i.e. the special case of  $\mathcal{K}_{12}$  where  $U_1 = U_2 = \mathbb{R}$ ), Eq. 26 reduces to:

$$\hat{\mathcal{K}}_{11}(B) = \frac{1}{\lambda^2} \int \int \frac{\mathbf{1}_B(y - x) \mathbf{1}_W(x) \mathbf{1}_W(y)}{\nu(W \cap (W + x - y))} \phi(dx) \phi(dy). \quad (27)$$

Following [9], we can express  $\mathcal{K}_{11}$  in terms of the Palm distribution  $P_o$ :

$$\mathcal{K}_{11}(B) = (1/\lambda) \mathbb{E}_{P_o}[\phi(B)], \quad (28)$$

where  $P_o(Y) = \int g(x) \mathbf{1}_Y(\phi + x) \phi(dx) P(d\phi)$ , with  $g(\cdot) : \mathbb{R}^d \rightarrow \mathbb{R}^+$  an arbitrary non-negative measurable function integrating to 1 and  $Y \in \mathcal{M}$ . Similarly, following [18] (Eq. 2.19) we can express the reduced second moment measure  $\mathcal{K}$  as:

$$\begin{aligned} \mathcal{K}(B) &= (1/\lambda) \mathbb{E}_{P_o}[\phi(B) \setminus \{o\}] \\ &= (1/\lambda) \mathbb{E}_{P_o}[\phi(B)] - (1/\lambda) \mathbb{E}_{P_o}[\{o\}] \\ &= \mathcal{K}_{11}(B) - \mathcal{K}_{11}(\{o\}). \end{aligned} \quad (29)$$

The Ripley  $K$  function is defined in terms of the reduced second moment measure, giving:

$$\begin{aligned} K(r) &= \mathcal{K}(B(o, r)) \\ &= \mathcal{K}_{11}(B(o, r)) - \mathcal{K}_{11}(\{o\}). \end{aligned} \quad (30)$$

Hence, by applying Eq. 27 to each term in Eq. 30, we can form an unbiased estimator for  $K$  as discussed in the text:

$$\begin{aligned} \hat{K}(r) &= \hat{\mathcal{K}}_{11}(B(o, r)) - \hat{\mathcal{K}}_{11}(\{o\}) \\ &= \frac{1}{\lambda^2} \int \int \frac{\mathbf{1}_W(x) \mathbf{1}_W(y)}{\nu(W \cap (W + x - y))} [|x - y| \leq r] \phi(dx) \phi(dy) - C, \end{aligned} \quad (31)$$

where  $C = (\int_W \phi(\{x\}) \phi(dx)) / (\lambda^2 \nu(W))$ .

## Appendix B. EM Algorithm to fit Weighted Sum of Poisson Distributions

Following Eq. 17, we can define a weighted sum of Poisson distributions with components  $m = 1 \dots M$  having weights  $w_1 \dots w_M$  and means  $\alpha_1 \dots \alpha_M$  by the distribution:

$$P(x \in R) = \sum_{\mathbf{n} \in (\mathbb{N} \cup \{0\})^M} [(\sum_m w_m n_m) \in R] \cdot \prod_m \text{Poisson}(n_m | \alpha_m), \quad (32)$$

where  $[A]$  is the Iverson bracket, which is 1 when  $A$  is true and 0 otherwise. We can reexpress this distribution in a form involving latent variables  $Z_1 \dots Z_M$ :

$$\begin{aligned} Z_m &\sim \text{Poisson}(\cdot; \alpha_m) \\ X &= \sum_m w_m z_m, \end{aligned} \quad (33)$$

or equivalently:

$$\begin{aligned} P(x) &= \sum_{\mathbf{z}} P(\mathbf{z}) [x = \sum_m w_m z_m] \\ P(\mathbf{z}) &= \prod_m \text{Poisson}(z_m; \alpha_m) \end{aligned} \quad (34)$$

where  $\mathbf{z} = [z_1, z_2, \dots, z_M]$ .

The EM algorithm can be applied to fit distributions involving latent variables, and maximizes the log-likelihood (to a local optimum) by introducing an auxiliary distribution  $q$  over the latent variables, and alternately minimizing the KL-divergence between  $q(\mathbf{Z})$  and the true marginal distribution over latent variables  $p(\mathbf{Z})$  (E-step), and maximizing the following lower-bound on the log-likelihood (M-step) (see [19,30]):

$$\mathcal{L}(q, \theta) = \sum_{\mathbf{Z}} q(\mathbf{Z}) \log \left( \frac{p(\mathbf{X}, \mathbf{Z} | \theta)}{q(\mathbf{Z})} \right), \quad (35)$$

where  $\theta$  are the distribution parameters, and  $\mathbf{X}, \mathbf{Z}$  are the observed data and a fixed setting of the latent variables respectively. We will write  $\mathbf{x}_i$  for the  $i$ 'th data point, and  $\mathbf{z}_{i,m}$  for the  $m$ 'th latent variable associated with the  $i$ 'th data point, for data points  $i = 1 \dots N$ . For the weighted sum of Poisson distributions, we assume that the weights  $w_1 \dots w_M$  are fixed in advance, and hence we optimize  $\theta = \{\alpha_1 \dots \alpha_m\}$ . Further, since the distribution only places positive probability on values which can be expressed in the form  $\sum_m w_m n_m$  for  $n_m \in \mathbb{N} \cup \{0\}$ , we assume that the values  $x_i$  have been rounded to the nearest such value.

**E-step.** The KL-divergence between  $q(\mathbf{Z})$  and  $p(\mathbf{Z})$  can be minimized by calculating the posterior distribution of  $\mathbf{z}_i$  for each data point given the current parameter settings. Writing  $\gamma_{i,\mathbf{n}}$  for the posterior  $P(\mathbf{z}_i = \mathbf{n} | \alpha'_1 \dots \alpha'_M)$ , where  $\alpha'_m$  is the current value of  $\alpha_m$ , these can be calculated as:

$$\gamma(i, \mathbf{n}) = \frac{[x_i = \sum_m w_m n_m] \prod_m \text{Poisson}(n_m; \alpha'_m)}{Z}, \quad (36)$$

where  $Z$  is a normalizing constant. Since for each data point there are only finitely many latent-variable settings for which  $\gamma(i, \mathbf{n})$  is non-zero (since any for which  $w_m n_m > x_i$  will be zero), the values  $\gamma(i, \mathbf{n})$  can be found for all  $\mathbf{n}$  by explicit calculation. For initialization, we set  $\gamma(i, \mathbf{n}) = 1$  for an arbitrary  $\mathbf{n}$  such that  $[x_i = \sum_m w_m n_m]$ .  $q$  is found by setting  $q(\mathbf{Z}) = \prod_i \gamma(i, \mathbf{z}_i)$ .

**M-step.** Substituting Eqs.34 and 36 into Eq. 35 yields:

$$\begin{aligned} \mathcal{L}(q, \alpha') &= K + \sum_{i,m,\mathbf{n}} \gamma(i, \mathbf{n}) \log(\text{Poisson}(n_m | \alpha'_m)) \\ &= K + \sum_{i,m,\mathbf{n}} \gamma(i, \mathbf{n}) \log \left( \frac{(\alpha'_m)^{n_m}}{n_m!} \exp(-\alpha'_m) \right), \end{aligned} \quad (37)$$

where  $K$  is the entropy of  $q$ . Eq. 37 can be seen to break into separate collections of terms involving each of the  $\alpha'_m$ 's. Differentiating with respect to  $\alpha'_m$  and setting to zero yields the update:

$$\alpha'_m = \frac{\sum_{i,\mathbf{n}} \gamma(i, \mathbf{n}) n_m}{N}. \quad (38)$$

## Appendix C. Experimental Methods for Fluorescent Microscopy Data

NIH/3T3 mouse fibroblast cells were serum-starved for 16 hours prior to seeding on Fibronectin crossbow micropatterned surfaces (individual micropatterns approximately  $25\mu\text{m}$  in height and width). The cells

were allowed to grow for various lengths of time (2, 3, 4, 5 and 7 hours) before fixing in formaldehyde, permeabilization, and hybridization of probes. RNA FISH probes were designed and applied using the method of [31], which targets multiple 20-mer oligonucleotide probes to each mRNA. Rabbit polyclonal anti-Arghdia, anti-Gapdh, anti- $\beta$ -Actin and anti-Par3 antibodies were used for the IF staining (Santa Cruz and Abcam). Rat monoclonal anti-tubulin antibody (Abcam) was used for tubulin staining in all cells, along with DAPI for nuclear staining.

Images were captured on a spinning disk confocal Revolution XD system (Andor). Each cell was imaged as an individual  $z$ -stack, with each image comprising  $512 \times 512$  pixels, 15-25  $z$ -levels, and approximately  $0.1 \mu\text{m}$  pixel width and  $0.3 \mu\text{m}$  separation between  $z$ -levels. Background subtraction was applied to all images using ImageJ (IF and FISH), and spot detection was performed to determine mRNA positions from the FISH  $z$ -stacks using [32].

2D segmentation of the nucleus region was performed by max-projecting the DAPI  $z$ -stacks, thresholding the resulting images, and applying image dilation to the binary masks. 2D segmentation of cellular regions was performed similarly by max-projection, thresholding and dilating the tubulin IF  $z$ -stacks. To estimate a height map across the cellular region (to construct a 3D cellular model), we first estimated the base  $z$ -level of the cell to be the level with the maximum total tubulin intensity (cells adhere to micropatterned regions on a 2D surface, and thus achieve greatest spread at their base). We then search at each 2D location for the  $z$ -level with the max-tubulin intensity above the base level, which we observed empirically to provide a reliable indicator of the cell boundary. The final height-map was formed by smoothing the resulting surface with a  $3 \times 3$  box filter.

## References

1. C. Lee, H. Zhang, A. E. Baker, P. Occhipinti, M. E. Borsuk, and A. S. Gladfelter. Protein Aggregation Behavior Regulates Cyclin Transcript Localization and Cell-Cycle Control. *Developmental Cell*, 25:572-584, 2013.
2. Stoyan, D. and Penttinen, A. Recent applications of point process methods in forestry statistics. *Statistical Science*, 61-78, 2000.
3. Gatrell AC, Bailey TC, Diggle PJ and Rowlingson BS. Spatial point pattern analysis and its application in geographical epidemiology. *Transactions of the Institute of British Geographers* 21: 256274, 1996.
4. B. D. Ripley. Modelling Spatial Patterns. *Journal of the Royal Statistical Society, Series B*, 39(2):172-212, 1977.
5. Lagache, T., Lang, G., Sauvonnet, N. and Olivo-Marin, J. C. Analysis of the spatial organization of molecules with robust statistics. *PLOS ONE*, 2013.
6. S. N. Chui, D. Stoyan, W. S. Kendall, and J. Mecke. *Stochastic Geometry and its Applications, Third Edition*, Wiley, 2013.
7. J. F. C. Kingman. Completely random measures. *Pacific Journal of Mathematics*, 21(1):5978, 1967.
8. M. I. Jordan. Hierarchical models, nested models and completely random measures. in *Frontiers of Statistical Decision Making and Bayesian Analysis: in Honor of James O. Berger*, New York: Springer, 2010.
9. D. Stoyan and J. Ohser. Cross-Correlation Measures of Weighted Random Measures and their Estimation. *Theory of Probability and its Applications*, 29(2):338-347, 1984.
10. N. Crosetto, M. Bienko, and A. van Oudenaarden. Spatially resolved transcriptomics and beyond. *Nature Reviews Genetics* 16/1: 57-66, 2015.
11. Blei, D. M., Griffiths, T. L. and Jordan, M. I. The nested Chinese restaurant process and Bayesian nonparametric inference of topic hierarchies. *Journal of the ACM (JACM)*, 57(2), 7, 2010.
12. Lin, D., Grimson, E. and Fisher III, J. W. Construction of dependent Dirichlet processes based on Poisson processes. In *Advances in neural information processing systems*, 2010.
13. Rao, V., and Teh, Y. W. Spatial normalized gamma processes. In *Advances in neural information processing systems*, 2009.
14. Sudderth, E. B. and Jordan, M. I. Shared segmentation of natural scenes using dependent Pitman-Yor processes. *Advances in Neural Information Processing Systems*, 2009.
15. J. F. C. Kingman. *Poisson Processes*, Oxford University Press, 2002.
16. J. Osher, and D. Stoyan. On the Second-Order and Orientation Analysis of Planer Stationary Point Processes. *Biometric Journal*, 23:523-533, 1981.
17. Davison, A. C. and Hinkley, D. V. *Bootstrap Methods and their Application*. Cambridge University Press, 1997.
18. Benes, V. and Rataj, J. *Stochastic Geometry: Selected Topics*. Springer Science and Buisness Media, 2007.
19. C. M. Bishop. *Pattern recognition and machine learning*. Springer, 2006.
20. K. Schauer, T. Duong, K. Bleakley, S. Bardin, M. Bornens, and B. Goud. Probabilistic density maps to study global endomembrane organization. *Nature Methods* 7:560 (2010).
21. Thery, M., Racine, V., Piel, M., Pepin, A., Dimitrov, A., Chen, Y., Jean-Baptiste, S. and Bornens, M. Anisotropy of cell adhesive microenvironment governs cell internal organization and orientation of polarity. *Proceedings of the National Academy of Sciences*, 2006.



22. S. Mili, K. Moissoglu, and I. G. Macara. Genome-Wide Screen Identifies Localized RNAs Anchored At Cell Protrusions Through Microtubules And APC. *Nature*, 453(7191):115, 2008.
23. J. Schmoranz, J. P. Fawcett, M. Segura, S. Tan, R. B. Vallee, T. Pawson, and G. G. Gundersen. Par3 and dynein associate to regulate local microtubule dynamics and centrosome orientation during migration. *Current Biology*, 19 (13), 1065-1074 (2009).
24. Buxbaum, A., Wu, B. and Singer, R.H. Single  $\beta$ -Actin mRNA Detection in Neurons Reveals a Mechanism for Regulating Its Translatability. *Science* 343:419, 2014.
25. E. Lécuyer, H. Yoshida, N. Parthasarathy, C. Alm, T. Babak, T. Cerovina, T. R. Hughes, P. Tomancak, and H. M. Krause. Global analysis of mRNA localization reveals a prominent role in organizing cellular architecture and function. *Cell* 131:174187, 2007.
26. Junker, J. P., Noël, E. S., Guryev, V., Peterson, K. A., Shah, G., Huiskens, J., McMahon AP, Berezikov E, Bakkers J, and van Oudenaarden, A. Genome-wide RNA tomography in the zebrafish embryo. *Cell*, 159:3, 662-675, 2014.
27. Lawrence, N. D., Sanguinetti, G., and Rattray, M. Modelling transcriptional regulation using Gaussian processes. In *Advances in Neural Information Processing Systems* 785-792, 2006.
28. Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D. A., and Nolan, G. P. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721), 523-529, 2005.
29. Snijder, B., Sacher, R., Rm, P., Damm, E. M., Liberali, P. and Pelkmans, L. Population context determines cell-to-cell variability in endocytosis and virus infection. *Nature*, 461(7263), 520-523, 2009
30. A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society B*. 1-38, 1977.
31. Raj, A., van den Boogard, P., Rifkin, S.A., van Oudenaarden, A. and Tyagi, S. (2008) Imaging individual mRNA molecules using multiple singly labeled probes. *Nature Methods* 877-9, 2008.
32. J. Olivo-Marin. Extraction of Spots in Biological Images using Multiscale Products. *Pattern Recognition*, 35(9):1989-1996, 2002.